Contents

[Are These Datasets Distinct? 3](#_Toc534132169)

[Option: ANOVA (Analysis of Variance) / F-Test 3](#_Toc534132170)

[Option T-Test, Welch’s T-Test / Welch’s ANOVA (New) 3](#_Toc534132171)

[Option: Z-Test. Large dataset, Central Limit Theorem, are the Averages different? 4](#_Toc534132172)

[Option: Student’s T-Test. The typical option. 4](#_Toc534132173)

[Option: Mann–Whitney *U Test* 4](#_Toc534132174)

[Option: K-S Test (if probabilities) 4](#_Toc534132175)

[Basic Questions: 4](#_Toc534132176)

[What is a Z-Test Statistic? 5](#_Toc534132177)

[What is TSS? What is RSS? ESS? 5](#_Toc534132178)

[The Total Sum of Squares 5](#_Toc534132179)

[Residual Sum of Squares 5](#_Toc534132180)

[Sum of Squares (Model) – Explained Sum of Squares 5](#_Toc534132181)

[Residual Standard Error 5](#_Toc534132182)

[Total Sum of Squares 5](#_Toc534132183)

[Explained Sum of Squares 6](#_Toc534132184)

[Residual Sum of Squares 6](#_Toc534132185)

[What is R^2? 6](#_Toc534132186)

[What is Adjusted R^2? 6](#_Toc534132187)

[What is an F-Test? (And ANOVA) 7](#_Toc534132188)

[What is a T-Test? 8](#_Toc534132189)

[What is K-S Test? 8](#_Toc534132190)

[What’s the Bias-Variance Tradeoff? 8](#_Toc534132191)

[What is p-value 9](#_Toc534132192)

[What is Statistical Power? – What is it Really? 10](#_Toc534132193)

[p-value Misinterpretations 10](#_Toc534132194)

[Goof 1: **The *P* value is the probability that the test hypothesis is true** 10](#_Toc534132195)

[What is lasso? 10](#_Toc534132196)

[What is cross-validation? 11](#_Toc534132197)

[What is bootstrap? 11](#_Toc534132198)

[Highly Unbalanced Datasets – How to Handle / Process? 11](#_Toc534132199)

[Basic Probability 11](#_Toc534132200)

[Standard Error 11](#_Toc534132201)

[Margin of Error 11](#_Toc534132202)

[Bayes Theorem 11](#_Toc534132203)

[Central Limit Theorem, Normal Distribution, Z-Test 12](#_Toc534132204)

[Z Score 12](#_Toc534132205)

[Basic 95% Confidence Interval (two-tailed) 12](#_Toc534132206)

[Type 1 vs Type 2 12](#_Toc534132207)

[Two Tailed 13](#_Toc534132208)

[Two Tailed vs One Sided – Be Careful 13](#_Toc534132209)

[ANOVA, K-S Test 13](#_Toc534132210)

[Linear Regression, How it Works, Weaknesses, Strengths, Results Review 13](#_Toc534132211)

[Understanding Step by Step 13](#_Toc534132212)

[The Total Sum of Squares 14](#_Toc534132213)

[Residual Sum of Squares 14](#_Toc534132214)

[Sum of Squares (Model) – Explained Sum of Squares 14](#_Toc534132215)

[Residual Standard Error 14](#_Toc534132216)

[R^2 - Proportion of Variation in Y that can be explained using X 14](#_Toc534132217)

[The F-Value 14](#_Toc534132218)

[Linear Model in R – Understanding the Output 15](#_Toc534132219)

[Least Squares Finds the Best Fit (Common) 16](#_Toc534132220)

[Issues with Linear 16](#_Toc534132221)

[Testing Models (e.g. Regression): Predicted vs Actual 17](#_Toc534132222)

[MSE: Mean Squared Error (Testing Regression) 17](#_Toc534132223)

[RSE: Residual Standard Error 17](#_Toc534132224)

[RMSE: Root Mean Squared Error 17](#_Toc534132225)

[MAE: Mean Absolute Error 17](#_Toc534132226)

[K-S Test 17](#_Toc534132227)

[Bias-Variance Tradeoff 17](#_Toc534132228)

[Cross Validation – (Expand Training Set) 18](#_Toc534132229)

[Statistical Power in Detail 18](#_Toc534132230)

[Bootstrapping 19](#_Toc534132231)

[Classifier: Support Vector Machines 19](#_Toc534132232)

[Some Basics / Terms and Purposes 19](#_Toc534132233)

[Accuracy, Precision, Recall, Confusion Matrix 19](#_Toc534132234)

[Accuracy: % of Population that’s properly classified 19](#_Toc534132235)

[Precision: % of Positives that are actually positive 19](#_Toc534132236)

[False Discovery Rate: % of Predictions that are False Positive 19](#_Toc534132237)

[Recall: % of Positives that are flagged as positive 20](#_Toc534132238)

[The Confusion Matrix 20](#_Toc534132239)

[ROC Curve 20](#_Toc534132240)

[Chi-Squared Test (vs Z-Test) 20](#_Toc534132241)

[1. A test of independence - determine if two variables are independent 20](#_Toc534132242)

[2. A Test of Fit – (Pearson’s) Do Categories Fit a (particular) distribution? 20](#_Toc534132243)

[Z-Test vs T-Test vs Chi Squared vs Fisher’s 20](#_Toc534132244)

[What are Degrees of Freedom? – Values In the Math That Can Vary 21](#_Toc534132245)

**Question**

# Are These Datasets Distinct?

### What’s still Needed:

Understanding the situation… ‘datasets distinct’ is now striking me as insufficient.

### Option: ANOVA (Analysis of Variance) / F-Test

* 1. F-Test. Variation between sample means / Variation within samples
  2. ANOVA uses the F-test to determine whether the variability between group means is larger than the variability of the observations within the groups.
     1. If that ratio is sufficiently large, you can conclude that not all the means are equal.
     2. You take many different samples and plot the means.

### Option T-Test, Welch’s T-Test / Welch’s ANOVA (New)

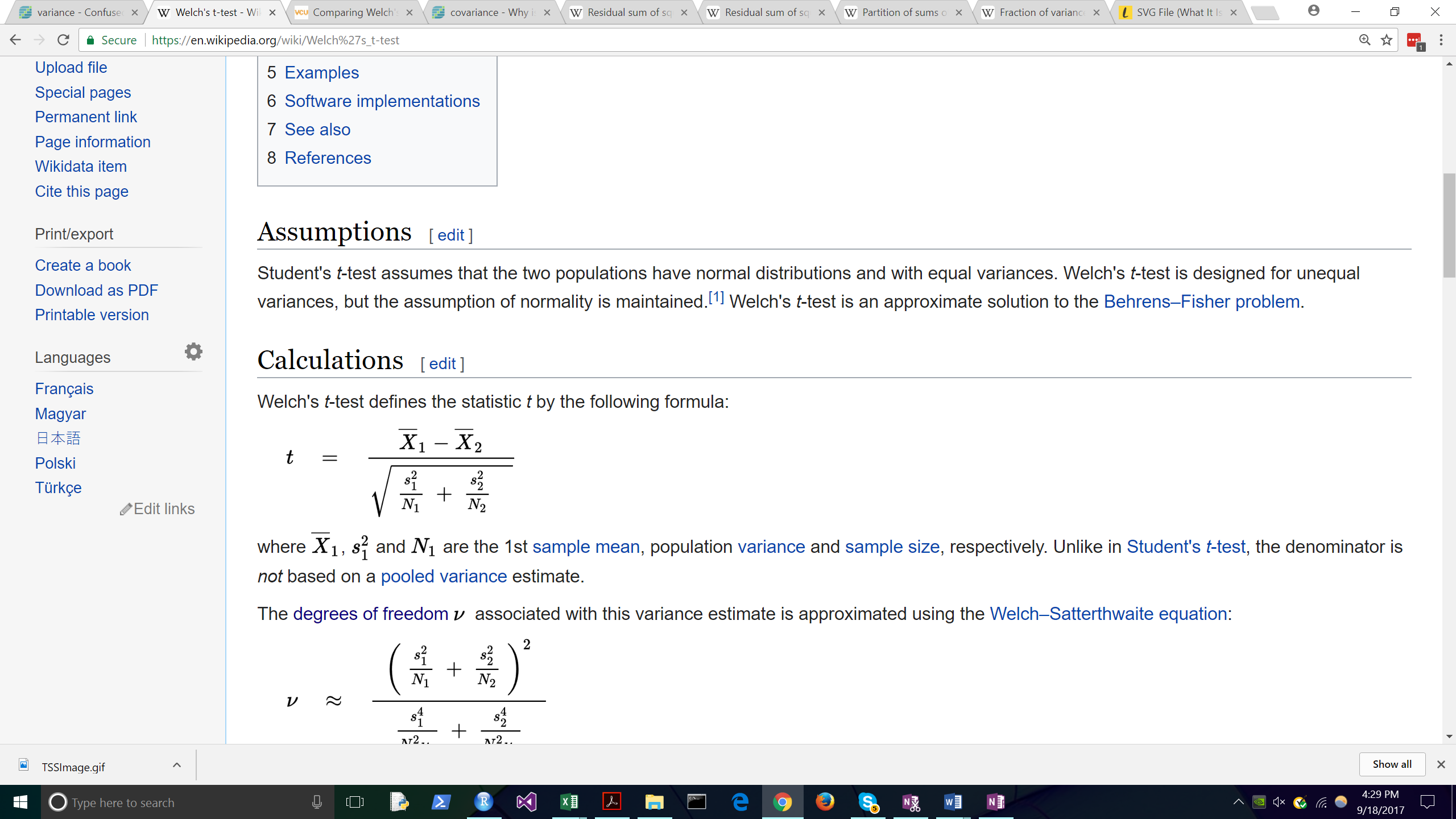
The T-Test is run when we don’t really know the standard deviation of our population. In other words – we can’t run population mean / population standard deviation vs sample mean / sample standard deviation. You instead base this on a flatter distribution.

Welch’s T-Test is used when variances between are thought NOT to be equal to each other, doing this by adjusting the degrees of freedom. It is now used BY DEFAULT when you run ‘T.test’ in R.

* In R it can be shut off by using the var.equal=TRUE argument.

<http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5026&context=etd>

Welch's *t*-test defines the statistic *t* by the following formula



Mean(X1) = the 1st sample mean, S1 = population variance N1 = sample size. Unlike in Student's t-test, the denominator is not based on a pooled variance estimate.

The degrees of freedom associated with this variance estimate is approximated using the Welch–Satterthwaite equation.

The Wikipedia Page (And Web Pages) I’ve read are VERY Welch’s t-test favorable. I’m somewhat suspicious, BUT – stats research DOES proceed.

* Welch's t-test is more robust than Student's t-test and maintains type I error rates close to nominal for unequal variances and for unequal sample sizes under normality.
* Furthermore, the power of Welch's t-test comes close to that of Student's t-test, even when the population variances are equal and sample sizes are balanced.
* Welch's t-test can be generalized to more than 2-samples, which is more robust than one-way analysis of variance (ANOVA)
* Welch's *t*-test can be applied directly and without any substantial disadvantages to Student's *t*-test as noted above. Welch's *t*-test remains robust for skewed distributions and large sample sizes

Welch’s T-Test in R: t.test(data1, data2, alternative="two.sided", var.equal=FALSE)

### Option: Z-Test. Large dataset, Central Limit Theorem, are the Averages different?

* 1. Normal Distribution

### Option: Student’s T-Test. The typical option.

* 1. Students T-distribution – wider, flatter than normal distribution

### Option: Mann–Whitney *U Test*

* 1. THIS I need to read up on further. It competes with the T-Test in some applications, and is apparently widely used – When Data Is not Normal.
  2. Requires: Dependent Variable is ordinal (e.g. 1-7) or continuous.
  3. Requires: Independent Variable is Binary category. (e.g. Male or Female)
  4. Requires: Independent Observations
  5. Requires: Distributions of your binary results are the same shape.

### Option: K-S Test (if probabilities)

* 1. Only for single variable distributions – of probability
  2. Compares the function of the sample to another function.
  3. Plot the two distributions.

# Basic Questions:

## What is a Z-Test Statistic?

It’s the genuine area under the curve on the ***normal distribution*.** You’re capturing 95% of the AUC of the group between -1.96 and 1.96. Basically 2 standard deviations (errors)

Here’s the basic formula: Z = (Sample Mean) – (Actual Mean) / (St Dev)

The Basic Assumption: you are working with values that are normally distributed. And the central limit theorem says that the averages of multiple samples will be normally distributed.

## What is R^2?

In Linear Regression, R^2 Represents the *proportion of variability in Y that can be explained using X.*

Formula: R^2 = (TSS *–* RSS) / TSS = 1*−*(RSS/TSS)

An *R*2 statistic that is close to 1 indicates that a large proportion of the variability in the response

has been explained by the regression.

A number near 0 indicates that the regression did not explain much of the variability in the response;

NOTE PLEASE: A VERY high R^2, e.g. 95% (.95), is often a red flag! A sign of overfitting!

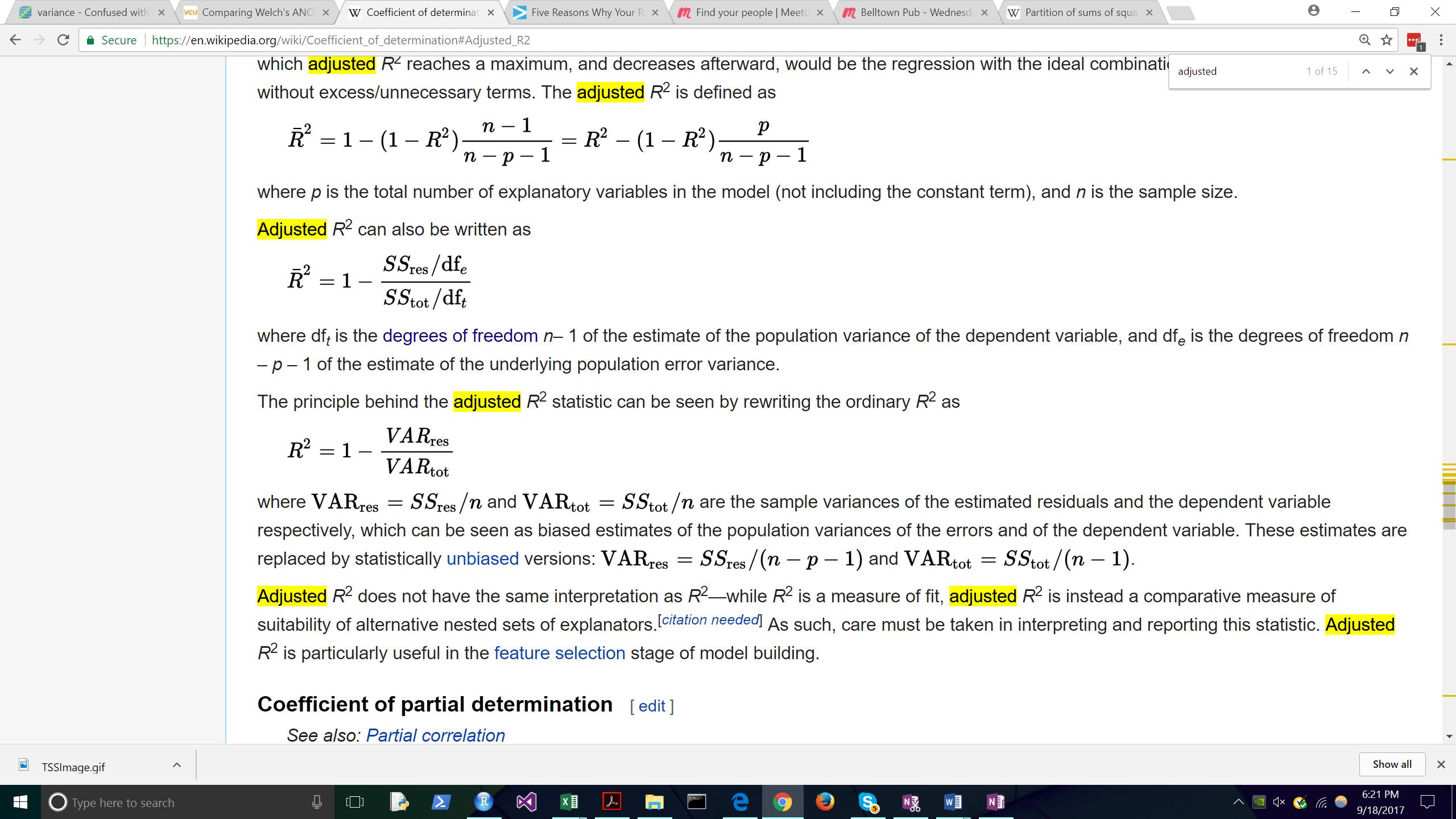
Also! Note that R^2 is only an estimate of the population!

## What is Adjusted R^2?

Adjusted R^2 is an adjustment to R^2, one based on the number of predictors in the model.

* A high Adjusted R^2 is a sign that the predictor improves the model greater than by chance could predict.

Adjusted *R*2 can be written as



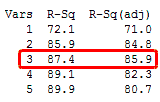
SSres = RSS (Residual Sum of Squares)

SStot = Total Sum of Squares.

Dft = degrees of freedom n-1 of the estimated population variance of the dependent variable.

Dfe = degrees of freedom n – p – 1 of the estimated underlying population variance.

In the below chart (scraped from a web page), the R^2 value keeps increasing as you add predictors, BUT the adjusted R^2 drops. An over-specified model can reduce precision.



USE:

Adjusted *R*2 is a comparative measure of suitability of alternative nested sets of explanatory variables. It’s most useful *during feature selection.*

Wikipedia:

The use of an adjusted *R*2 is an attempt to take account of the phenomenon of the *R*2 automatically and spuriously increasing when extra explanatory variables are added to the mode.

If a set of explanatory variables with a predetermined hierarchy of importance are introduced into a regression one at a time, with the adjusted *R*2 computed each time, the level at which adjusted *R*2 reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess/unnecessary terms

## What is an F-Test? (And ANOVA)

F = ((TSS − RSS)/p)/ (RSS/(n − p − 1))

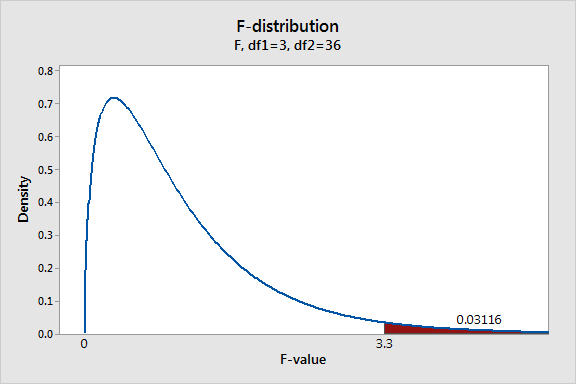
ANOVA uses the F-test to determine whether the variability between group means is larger than the variability of the observations within the groups. If that ratio is sufficiently large, you can conclude that not all the means are equal.

In One Way Anova the ratio of the between-group variability to within-group variability follows an F-distribution when the null hypothesis is true.

When you perform a one-way ANOVA for a single study, you obtain a single F-value. However, if we drew multiple random samples of the same size from the same population and performed the same one-way ANOVA, we would obtain many F-values and we could plot a distribution of all of them.

The F-distribution assumes that the null hypothesis is true, we can place the F-value from our study in the F-distribution to determine how consistent our results are with the null hypothesis and to calculate probabilities.

This probability that we’re calculating is also known as the p-value!



RSS: Residual Sum of Squares

* In linear regression, the attempt is to minimize the RSS stat.
* It’s the variances of the values from a straight line.

TSS: Total Sum of Squares

* TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
  + We are saying that the sample is going to have variance already.

In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.

## What is a T-Test?

* 1. **The student’s t test. We use a flatter distribution than the normal to establish confidence intervals – if the value is outside those intervals, the null hypothesis (things are the same) is rejected.**
  2. **Welch’s T Test.**

## What is K-S Test?

## What’s the Bias-Variance Tradeoff?

A good summation & more commentary is found here:

<https://insidebigdata.com/2014/10/22/ask-data-scientist-bias-vs-variance-tradeoff/>

The error due to variance is the amount by which the prediction, over one training set, differs from the expected predicted value, over all the training sets.

The error due to squared bias is the amount by which the expected model prediction differs from the true value or target, over the training data.

Models that exhibit small variance and high bias underfit the truth target

Models that exhibit high variance and low bias overfit the truth target.

The “tradeoff” between bias and variance can be viewed in this manner – a learning algorithm with low bias must be “flexible” so that it can fit the data well. But if the learning algorithm is too flexible (for instance, too linear), it will fit each training data set differently, and hence have high variance.

## What is p-value

LOTS OF NOTES CAN BE PUT HERE – PROPER UNDERSTANDING OF P-VALUES

Some notes from this publication:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/>

Not only does a P value not tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct—an assurance that is lacking in far too many studies.

The P value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility.

Too often, however, the P value is degraded into a dichotomy in which results are declared “statistically significant” if P falls on or below a cut-off (usually 0.05) and declared “nonsignificant” otherwise.

*The*P*value is a probability computed*assuming*chance was operating alone.*

A small P value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct;

it may be small because there was a large random error or because some assumption other than the test hypothesis was violated

A P ≤ 0.05 only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large or larger than that observed no more than 5 % of the time if only chance were creating the discrepancy. --- RIGHT! THIS I have long understood. Chance could have the null hypothesis true AND produce that result --- 5 times out of 100 tries.

many authors will misinterpret P = 0.70 from a test of the null hypothesis as evidence for no effect, when in fact it indicates that, even though the null hypothesis is compatible with the data under the assumptions used to compute the P value, it is not the hypothesis most compatible with the data—that honor would belong to a hypothesis with P = 1

## What is Statistical Power? – What is it Really?

* In Brief: The probability of staying with H0 when in fact it is false.

## What is lasso?

## What is cross-validation?

A method of doing both train and test on the same dataset separately. This essentially expands the amount of data available for training with.

It’s accomplished by splitting the dataset into sections, e.g. 10 folds. In the first set, you train on 9 and test on the 10th. In the second, you choose a different fold as the test set.

It’s not a perfect solution – after all, if your small dataset is under-representing an important characteristic of a major population segment, then looking at it from 10 different angles still won’t produce an algorithm to target said segment!

But it’s a darn sight better than splitting your small dataset widely! Very commonly used.

## What is bootstrapping?

The basic idea of bootstrapping is that inference about a population from sample data, (sample → population), can be modeled by *resampling* the sample data and performing inference about a sample from resampled data, (resampled → sample).

It doesn’t create more information.

# Misinterpretations

Much of these notes are resourced from this publication.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/>

### p-value Goof : **The P value is the probability that the test hypothesis is true**

NO: The Pvalue assumes the test hypothesis is true—it is not a hypothesis probability.

The P value simply indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the test (the underlying statistical model).

Thus P = 0.01 would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while P = 0.40 would indicate that the data are much closer to the model prediction, allowing for chance variation.

### Power Goofs Misinterpretations

Goof: If accept null, and power of your test is 90% the chance you are in error (false negative) is 10%

NO. If the null hypothesis is false and you accept it, the chance you are in error is 100 %.

The 10 % refers only to how often you would be in error over very many uses of the test across different studies when the particular alternative used to compute power is correct and all other assumptions used for the test are correct in all the studies.

Despite its shortcomings for interpreting current data, power can be useful for designing studies and for understanding why replication of “statistical significance” will often fail even under ideal conditions.

Studies are often designed or claimed to have 80 % power against a key alternative when using a 0.05 significance level, although in execution often have less power due to unanticipated problems such as low subject recruitment.

# Highly Unbalanced Datasets – How to Handle / Process?

1. We have undersampling of the majority and oversampling of the minority as techniques available. BUT there are considerations.
2. You CAN train using over-sampled, but you cannot TEST using the oversampled dataset.
   1. Your test must be representative of reality.
3. On some occasions a stratified sample can work – set up your datasets such that each train and test has the same proportion of positive-labeled points. But… you lose some random error guarantees.

# Bayesian Statistics

## Book A

We Use Prior Probabilities to Predict New Ones

The *likelihood* associated with a particular model is the probability of the observed data under that model

*P*(*Model | Data*) ∝*P*(*Model*)*×P*(*Data | Model*)

## Book B

The major input of the Bayesian approach, compared with a traditional

likelihood approach, is that it modifies the likelihood function into a *posterior*

distribution, which is a valid probability distribution on *Θ* defined by the

classical Bayes’ formula (or theorem)

the prior distribution summarizes the *prior information* on *θ*; that is, the knowledge that is available on *θ prior* to the observation of the sample *Dn*.

Bayes provides a fully probabilistic framework for the inferential analysis, with respect to a reference measure *π*(*θ*).

### Book C

Suppose that we wish to classify an observation into one of *K* classes. Let’s choose a response variable *Y – which* can take on *K* possible distinct and unordered values. Let *πk* represent the overall or *prior* probability that a randomly chosen observation comes from the *k*th class;

This is the probability that a given observation is associated with the *k*th category of the response variable *Y* .

Let *fk*(*X*) *≡* Pr(*X* = *x|Y* = *k*) denote the *density function* of *X* for an observation that comes from the *k*th class

In other words, *fk*(*x*) is relatively large if there is a high probability that an observation in the *k*th class has *X ≈ x*, and *fk*(*x*) is small if it is very unlikely that an observation in the *k*th class has *X ≈ x*.

Then *Bayes’ theorem* states that

The god damn book starts putting together a formula with at least one of the values being not fully defined. God damn it.

### Wikipedia

<https://simple.wikipedia.org/wiki/Bayes%27_theorem>

A simple example is as follows: There is a 40% chance of it raining on Sunday. If it rains on Sunday, there is a 10% chance it will rain on Monday. If it didn't rain on Sunday, there's an 80% chance it will rain on Monday.

"Raining on Sunday" is event A, and "Raining on Monday" is event B.

* P( *A* ) = 0.40 = Probability of Raining on Sunday.
* P( *A`* ) = 0.60 = Probability of not raining on Sunday.
* P( *B | A* ) = 0.10 = Probability of it raining on Monday, if it rained on Sunday.
* P( *B` | A* ) = 0.90 = Probability of it not raining on Monday, if it rained on Sunday.
* P( *B | A`* ) = 0.80 = Probability of it raining on Monday, if it did not rain on Sunday.
* P( *B` |A`* ) = 0.20 = Probability of it not raining on Monday, if it did not rain on Sunday.

The first thing we'd normally calculate is the probability of it raining on Monday: This would be the sum of the probability of "Raining on Sunday and raining on Monday" and "Not raining on Sunday and raining on Monday"

{\displaystyle 0.40\times 0.10+0.60\times 0.80=0.52=52\%} chance

However, what if we said: "It rained on Monday. What is the probability it rained on Sunday?" That is where Bayes' theorem comes in. It allows us to calculate the probability of an earlier event, given the result of a later event.

The equation used is:

{\displaystyle P(A|B)={\frac {P(B|A)\,P(A)}{P(B)}}.}In our case, "Raining on Sunday" is event A, and "Raining on Monday" is event B.

* P(*B|A*) = 0.10 = Probability of it raining on Monday, if it rained on Sunday.
* P(*A*) = 0.40 = Probability of Raining on Sunday.
* P(*B*) = 0.52 = Probability of Raining on Monday.

So, to calculate the probability it rained on Sunday, given that it rained on Monday:

{\displaystyle P(A|B)={\frac {P(B|A)\,P(A)}{P(B)}}.}

{\displaystyle P(A|B)={\frac {0.10\*0.40}{0.52}}=.0769}In other words, if it rained on Monday, there's a 7.69% chance it rained on Sunday.

## **Intuitive explanation**

To calculate the probability of it having rained on Sunday, given that it rained on Monday, we can take the following steps:

* We know that it rained on Monday. Therefore, the total probability is P(B).
* The probability it rained on Sunday is P(A).
* The probability it rained on Monday, given that it rained on Sunday is P(B|A).
* The probability of raining on Sunday AND raining Monday is P(A)\*P(B|A).
* Therefore, the total probability of it having rained on Sunday, given that it rained on Monday, is the chance of it raining on Sunday and Monday divided by the total chance of it having rained on Monday.

Therefore,

{\displaystyle P(A|B)={\frac {P(B|A)\,P(A)}{P(B)}}.}Another way to see this, which shows where Bayes' theorem comes from, is to consider the probability P(AB) that it rains on both Sunday and Monday. This can be calculated in two different ways, which give the same answer for P(AB):

{\displaystyle P(A)\,P(B|A)=P(B)\,P(A|B)}Bayes' theorem is just another way to write that equation

## Bayes Theorem

**Probability (A|B) = (Probability (B|A)\*Probability(A))/ Probability (B)**

**a.k.a.**

**Probability of A, given B =**

**Probability of B existing with A \* Probability of A without B**

**Divided by Probability of B without A**

**Definitions**

* **A = A, without B**
* **B = B, without A**
* **A|B = (A), given (B being true)**
* **B|A = (B), given (A being true)**

Example: What’s the chance that any one person sampled will be flagged as a drug user, where the drug test is 99% accurate and the population is 5% actual?

* Produces a result of 33.2%. It’s more likely that the person does not use, when flagged as using.

An Example from ‘Applied Bayesian Statistics’:

We’re searching for a solution to a problem – we don’t always have repeatable experiments -

The probability of new data incoming is proportional to prior probability \* likelihood.

Standard Error **= (Sample SD) / SQRT(N)**

Margin of Error **= z\* (Standard Error)**

**Probability (A|B) = (Probability (B|A)\*Probability(A))/ Probability (B)**

**a.k.a.**

**Probability of A, given B =**

**Probability of B existing with A \* Probability of A without B**

**Divided by Probability of B without A**

**Definitions**

* **A = A, without B**
* **B = B, without A**
* **A|B = (A), given (B being true)**
* **B|A = (B), given (A being true)**

**The probability of new data incoming is proportional to prior probability \* likelihood.**

Example: What’s the chance that any one person sampled will be flagged as a drug user, where the drug test is 99% accurate and the population is 5% actual?

* Produces a result of 33.2%. It’s more likely that the person does not use, when flagged as using.

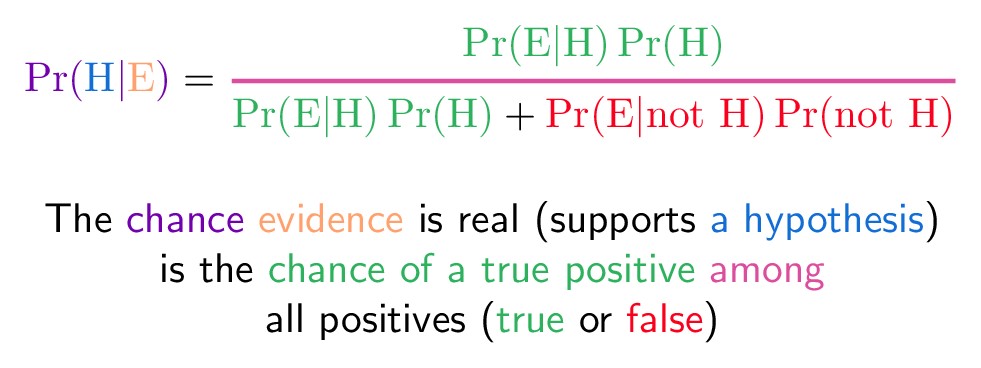
## Article from ‘BetterExplained’

<https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>

**Bayes’ theorem converts the results from your test into the real probability of the event.** For example, you can:

* **Correct for measurement errors**. If you know the real probabilities and the chance of a false positive and false negative, you can correct for measurement errors.
* **Relate the actual probability to the measured test probability.** Given mammogram test results and known error rates, you can predict the actual chance of having cancer given a positive test. In technical terms, you can find Pr(H|E), the chance that a hypothesis H is true given evidence E, starting from Pr(E|H), the chance that evidence appears when the hypothesis is true.

Bayes’ Theorem…. lets you take the test results and correct for the “skew” introduced by false positives. You get the real chance of having the event. Here’s the equation:



It all comes down to the chance of a *true* positive divided by the chance of *any*positive. We can simplify the equation to:

Pr(E) tells us the chance of getting *any* positive result, whether a true positive in the cancer population (1%) or a false positive in the non-cancer population (99%). In acts like a weighting factor, adjusting the odds towards the more likely outcome.

Forgetting to account for false positives is what makes the low 7.8% chance of cancer (given a positive test) seem counter-intuitive. Thank you, normalizing constant, for setting us straight!

Intuitive Understanding: Shine The Light

Consider a real population. You do some tests which “shines light” through that real population and creates some test results. If the light is completely accurate, the test probabilities and real probabilities match up. Everyone who tests positive is actually “positive”. Everyone who tests negative is actually “negative”.

But this is the real world. Tests go wrong. Sometimes the people who have cancer don’t show up in the tests, and the other way around.

Bayes’ Theorem lets us look at the skewed test results and correct for errors, recreating the original population and finding the real chance of a true positive result.

One clever application of Bayes’ Theorem is in [spam filtering](https://en.wikipedia.org/wiki/Bayesian_spam_filtering). We have

* Event A: The message is spam.
* Test X: The message contains certain words (X)

Plugged into a more readable formula (from Wikipedia):

Bayesian filtering allows us to predict the chance a message is really spam given the “test results” (the presence of certain words). Clearly, words like “viagra” have a higher chance of appearing in spam messages than in normal ones.

Spam filtering based on a blacklist is flawed — it’s too restrictive and false positives are too great. But Bayesian filtering gives us a middle ground — we use probabilities. As we analyze the words in a message, we can compute the chance it is spam (rather than making a yes/no decision). If a message has a 99.9% chance of being spam, it probably is. As the filter gets trained with more and more messages, it updates the probabilities that certain words lead to spam messages. Advanced Bayesian filters can examine multiple words in a row, as another data point.

# Central Limit Theorem, Normal Distribution, Z-Test

**Central Limit Theorem (Basically)**

1. **Upon many separate samples of a population, we will have many different means. However, those averages themselves – the distribution of those means is shaped normally.**
   * **THEREFORE We have strong reasons to believe that the ACTUAL population mean is somewhere within**
2. **Population Mean vs Sample Mean**
   1. **Sample Mean is random, distribution of sample means is normal.**

### Z Score

**It’s the genuine area under the curve on the normal distribution. You’re capturing 95% of the AUC of the group between -1.96 and 1.96. Basically 2 standard deviations (errors)**

**Here’s the basic formula: Z = (Sample Mean) – (Actual Mean) / (St Dev)**

* **Z = HA – H0/SE**
* **Then you look up the P-Value for that Z score. Think pnorm in R.**
* **A Very Low P Value (< .05), then we reject H0.**

### Basic 95% Confidence Interval (two-tailed)

**~1.96 \* StandardError = SampleMean – Actual Mean**

**SampleMean (+ or –) (1.96 \* StandardError) = Actual Mean**

### Type 1 vs Type 2

* **Type 1: Reject Null, Inappropriately (Narrow Scoop)**
* **Type 2: Accept Null, Inappropriately (Wide Scoop)**

### Two Tailed

* **Look Above or Below. P-Value is calc’d slightly different because you have two directions tested.**

### Two Tailed vs One Sided – Be Careful

**Significance vs Confidence Interval are Different with One Sided Tests.**

* Confidence interval and two sided hypothesis tests are functionally equivalent.
* But a single sided hypothesis test with .05% corresponds to a 90% interval. (.05 both sides)
* BE CAREFUL.

### ANOVA, K-S Test

1. **ANOVA**
2. **K-S Test to Test if Distributions are Similar. Can run 10000 ks tests across data, 10000 simulations, get a histogram, to find the probability that they are similar.**
3. **Shapiro-Wilk test can test for normality. (could look up statistical moments, to dive deeper). 3rd moment= skew, 4th moment = flatness**

# Linear Regression, How it Works, Weaknesses, Strengths, Results Review

**Coefficients (Parameters) = Intercept and Slope**

## Understanding Step by Step

**From an EXCELLENT post on the topic:**

[**https://stats.stackexchange.com/questions/256726/linear-regression-what-does-the-f-statistic-r-squared-and-residual-standard-err/256821#256821**](https://stats.stackexchange.com/questions/256726/linear-regression-what-does-the-f-statistic-r-squared-and-residual-standard-err/256821#256821)

The Process:

In a regression (or ANOVA), we build a model based on a sample dataset which enables us to predict outcomes from a population of interest.

To do so, the following three components are calculated in a simple linear regression from which the other components can be calculated:

* the mean squares (TSS)
* the F-value
* the R^2 (also the adjusted R^2)
* the residual standard error (RSE):

### The Total Sum of Squares -- Review

The TSS assess how well the mean fits the data. Why the mean?

Because the mean is the simplest model we can fit and hence serves as the model to which the least-squares regression line is compared to.

### Residual Sum of Squares - Review

The SS Residual (RSS) assess how well the regression line fits the data.

### Sum of Squares (Model) – Explained Sum of Squares

The SS (Model) compares how much better the regression line is compared to the mean (i.e. the difference between the TSS and the RSS).

### Residual Standard Error

The residual standard error (RSE) is the square root of the residual mean square. What’s this mean? Well – residual mean square = the averaged RSS.

Yes, RSE is a measurement of your observed data from the model. A low RSE is a good fit.

* the RSE tells you something about the inaccuracy of the model

### R^2 - Proportion of Variation in Y that can be explained using X

*R*2 = (TSS *–* RSS)/TSS = ESS/TSS

* R^2 is a statistic between 0 and 1.
* For a dataset that’s expected to be linear, we’d expect R^2 to be close to 1.
  + But others? – knowing the target is an art.

Rephrased: The R^2 tells you how much variation is explained by the model relative to the variation that was explained by the mean alone

### The F-Value

The F-value on the other is calculated as the model mean square MSmodel (or the signal) divided by the MSresidual (noise):

The F-value expresses how much of the model has improved (compared to the mean) given the inaccuracy of the model.

The F-statistic (in regression) is the division of the model mean square and the residual mean square. Software, after fitting a regression model, also provide the p-value associated with the F-statistic.

This allows you to test the null hypothesis that your model's coefficients are zero. You could think of it as the "statistical significance of the model as a whole."

**From One Way ANOVA: (F = variation between sample means / variation within the samples**

The F-statistic is the test statistic for F-tests

### Linear Model in R – Understanding the Output

lm(formula = dist ~ speed.c, data = cars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -29.069 -9.525 -2.272 9.215 43.201

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

## speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 15.38 on 48 degrees of freedom

## Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

## F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

In the Above, using the R ‘Cars’ dataset. There are 50 cars.

#### Residuals Section:

Residuals are the difference between the actual observed response values and the response values that the model predicted.

R gives them to us in 5 summary points. You should look for a symmetrical distribution across these points on the mean value zero (0). In the example above, the distribution of the residuals do not appear to be strongly symmetrical.

1. Residual Standard Error = 15.38 on 48 degrees of freedom.
   1. It’s an average of the residual standard error (regression line vs actual data.)
2. R-squared = .6511 and adjusted R-squared .6438.

#### Intercept

Here, it’s essentially the expected value, based on an average.

#### Estimate, Std Error

The slope of the regression line. In the example, you get 3.9324 in the Y-value for every 1 in X. The standard error is the standard error of this value if we computed many times.

#### T Value

Measures how many standard deviations the coefficient estimate is from 0. A large T-value means you’re likely to reject the null hypothesis. That you’ll find significance.

#### Coefficient - Pr(>t)

The p-value. .05 for 95% confidence on a normal distribution.

* The further from 0, the more likely you can reject the null hypothesis.

#### R^2

In this example, roughly 65% of the variance (or 64.38) can be explained by this one variable.

To take this cars dataset further – roughly 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed).

#### **F-Statistic**

Measures whether there’s a relationship between the predictor and the response variables. A high F-statistic is a good sign. F-statistic = Model’s Mean Square / Residual Mean Square

Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis. The reverse is true if the number of data points is small.

### Least Squares Finds the Best Fit (Common)

**Least Squares: It’s measuring the distance between the actual and the line.**

**Take the error (actual – prediction) across all points, and square them.**

**The line with the lowest total squared = the best fit.**

### Issues with Linear

1. Linear -- It’s a linear method, Data may not be linear.
2. Errors Not Constant – Least squares assumes that the variance for each coefficient’s error is static across the line. But this often isn’t the case in real life!
   1. Look at the Residual Plot. If it’s straight line, we’re good. If it’s funnel, we have heteroscedasticity. It’s imperfect.
   2. Option 1: Log-transforming the response (transform the predicted variable)
   3. Option 2: Do a weighted least squares calculation.
   4. Error Variances are Different - The least squared errors assumes that all variables have a constant variance. . It will never have a strong fit to a set of relationships that are not linear.
3. Outlier – the measured actual is far from the value predicted by the model.
   1. If it’s a measurement error, remove.
   2. If it’s not an error, perhaps the model needs adjusting, add a new predictor.
4. High Leverage – the predictor has an unusual value, capable of pulling the model.
5. Error Correlations - There may be correlations between the variables’ errors
   1. E.g. Time series data where two measurements are VERY close to each other.
6. Collinearity – close relationship (correlation) between predictors.
   1. Which results in a decline of the T-statistic. (Parameter Value / (Actual Val-Predicted Val) Divide parameter by standard error
   2. Correlation matrix can help but can be insufficient. Sometimes the correlation is multivariate.
   3. Compute the Variance Inflation Factor, a VIF > 5 or 10 indicates a problematic amount of multicollinearity.
      1. VIF(Coefficient) = 1/(1- (R squared of a regression between Coefficient vs other predictors/coefficients)

# Testing Results, Including Regression

In a Nutshell:

* ESS measures how much variation there is in the modelled values
* TSS measures how much variation there is in the observed data
* RSS measures the variation in modelling errors.

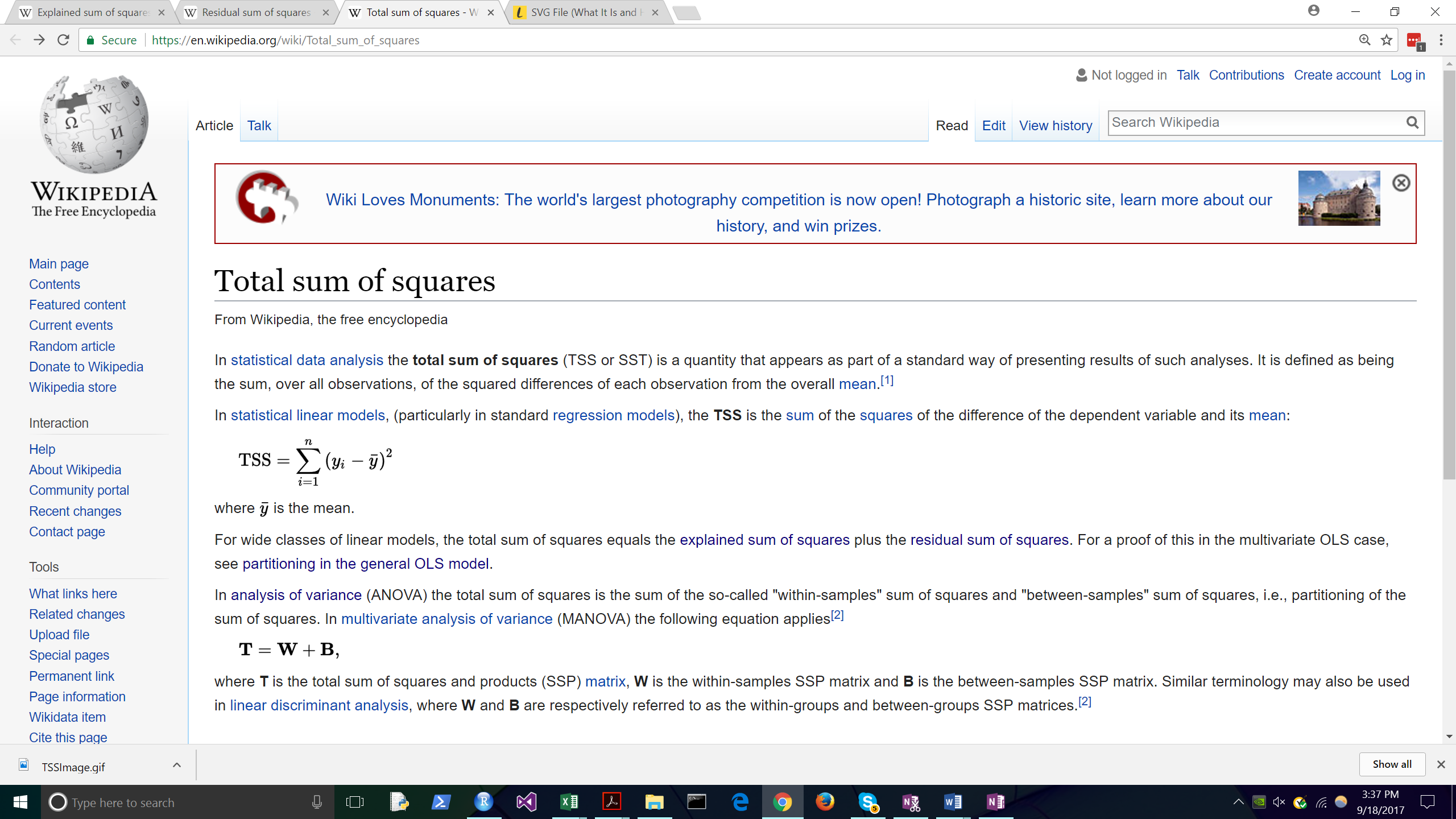
More detail given under ‘Linear Regression’

### The Total Sum of Squares

The TSS assess how well the mean fits the data. Why the mean?

Because the mean is the simplest model we can fit and hence serves as the model to which the least-squares regression line is compared to.

The **TSS (Total Sum of Squares)** is the sum of the squares of the difference of the dependent variable and its mean.



*yi* is the *i* th value of the variable to be predicted, *xi* is the *i* th value of the explanatory variable, and {\displaystyle f(x\_{i})}f(x) is the predicted value of *y.*

(Value – Predicted Value)^2

### Residual Sum of Squares – In Regression

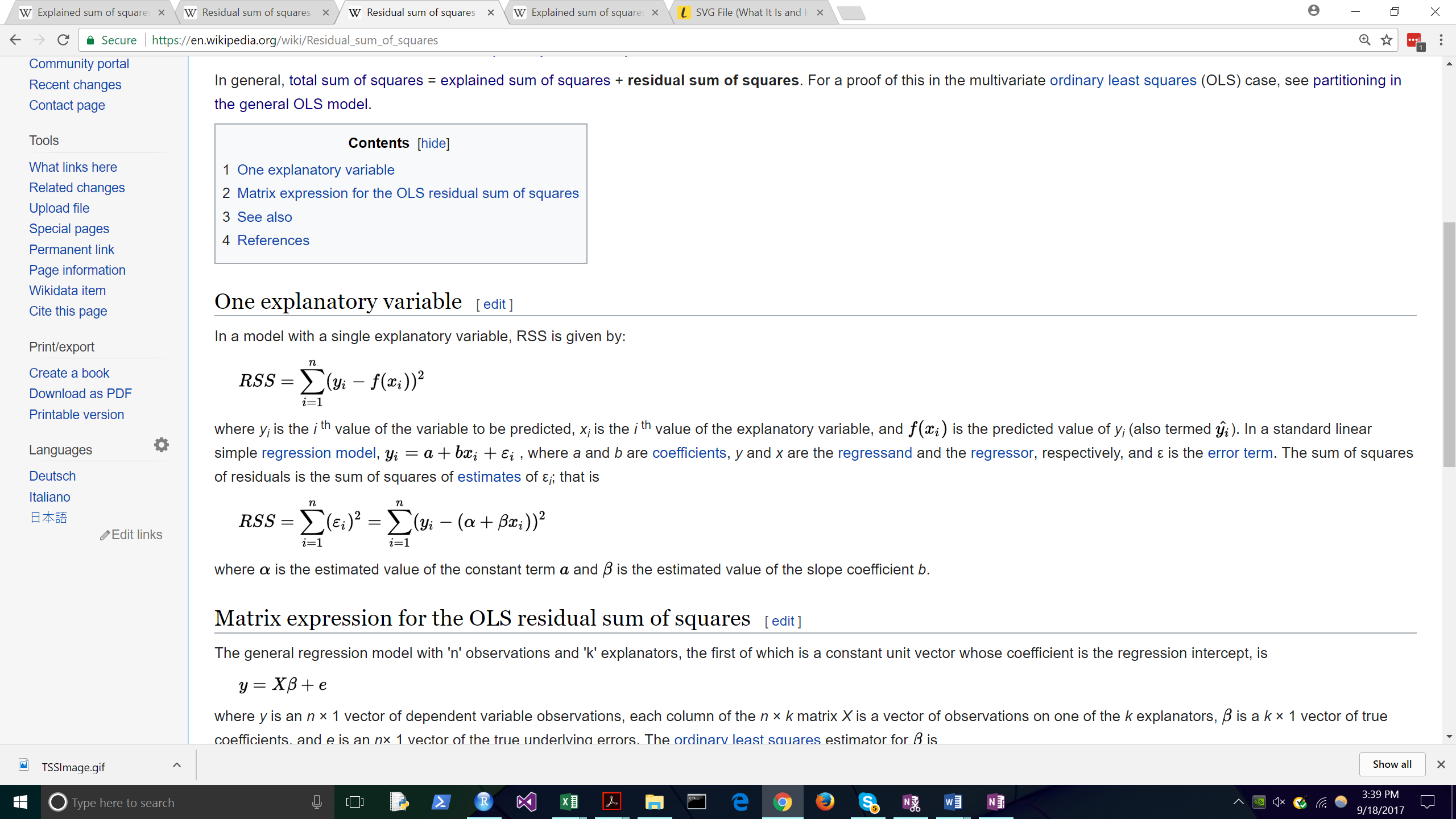
The SS Residual (RSS) assess how well the regression line fits the data.

*In Regression:* the mean is the simplest model we can fit and hence serves as the model to which the least-squares regression line is compared to.

In Regression it can also be thought of as Residual Sum of Squares + Explained Sum of Squares.

### Residual Sum of Squares

The RSS (Residual Sum of Squares) is the sum of the squares of residuals (deviations predicted from actual empirical values of data).



### Sum of Squares (Model) – Explained Sum of Squares

The SS (Model) compares how much better the regression line is compared to the mean (i.e. the difference between the TSS and the RSS).

### Residual Standard Error

The residual standard error (RSE) is the square root of the residual mean square. What’s this mean? Well – residual mean square = the averaged RSS.

Yes, RSE is a measurement of your observed data from the model. A low RSE is a good fit.

### Explained Sum of Squares

*The* **ESS** (explained sum of squares) is a quantity used in describing how well a model, often a regression model, represents the data being modelled.

It is the sum of the squares of the deviations of the predicted values from the mean value of a response variable.

(Predicted Value – Mean Value)^2

# Testing Models (e.g. Regression): Predicted vs Actual

### MSE: Mean Squared Error (Testing Regression)

For each new response / new value we have a series of coefficients, and a predicted result. The error is actual – predicted. Square that value, then take the average of it.

* MSE specifically highlights / bumps up those values that are farthest from predicted value. It’s small when the whole set is close to the predicted.

### RSE: Residual Standard Error

### RMSE: Root Mean Squared Error

RSME: Root of the Mean Squared Error

### MAE: Mean Absolute Error

mean absolute error (MAE) is a measure of difference between two continuous variables – that are paired. E.g. Predicted vs Actual

* Take absolute value of Y – X across the entire set, divide by n.
* It’s conceptually easy to understand.

### K-S Test

With the K-S Test we are testing our sample variables against either an existing distribution, or against each other.

**In practice, the statistic requires a relatively large number of data points to properly reject the null hypothesis.**

### Bias-Variance Tradeoff

Good test set performance of a statistical learning method requires low variance as well as low squared bias.

#### Error from Bias

Bias comes from the design of the model. The model is an attempt to match reality, but reality will be messy. E.g. pulling from a certain demographic can create bias.

#### Error from Variance

Variance comes from the model results – you might get a wide variance in the model results if you take many samples of small size. (e.g. 15 samples taken, sample size 10 each – the results will pop)

#### What’s the Tradeoff?

It’s about Overfitting. When you’ve got overfitting in your model, a model that is so tightly tied to your source dataset, you have low bias – but a strong potential for variance error in model results.

# Statistical Power in Detail

<http://rpsychologist.com/d3/NHST/>

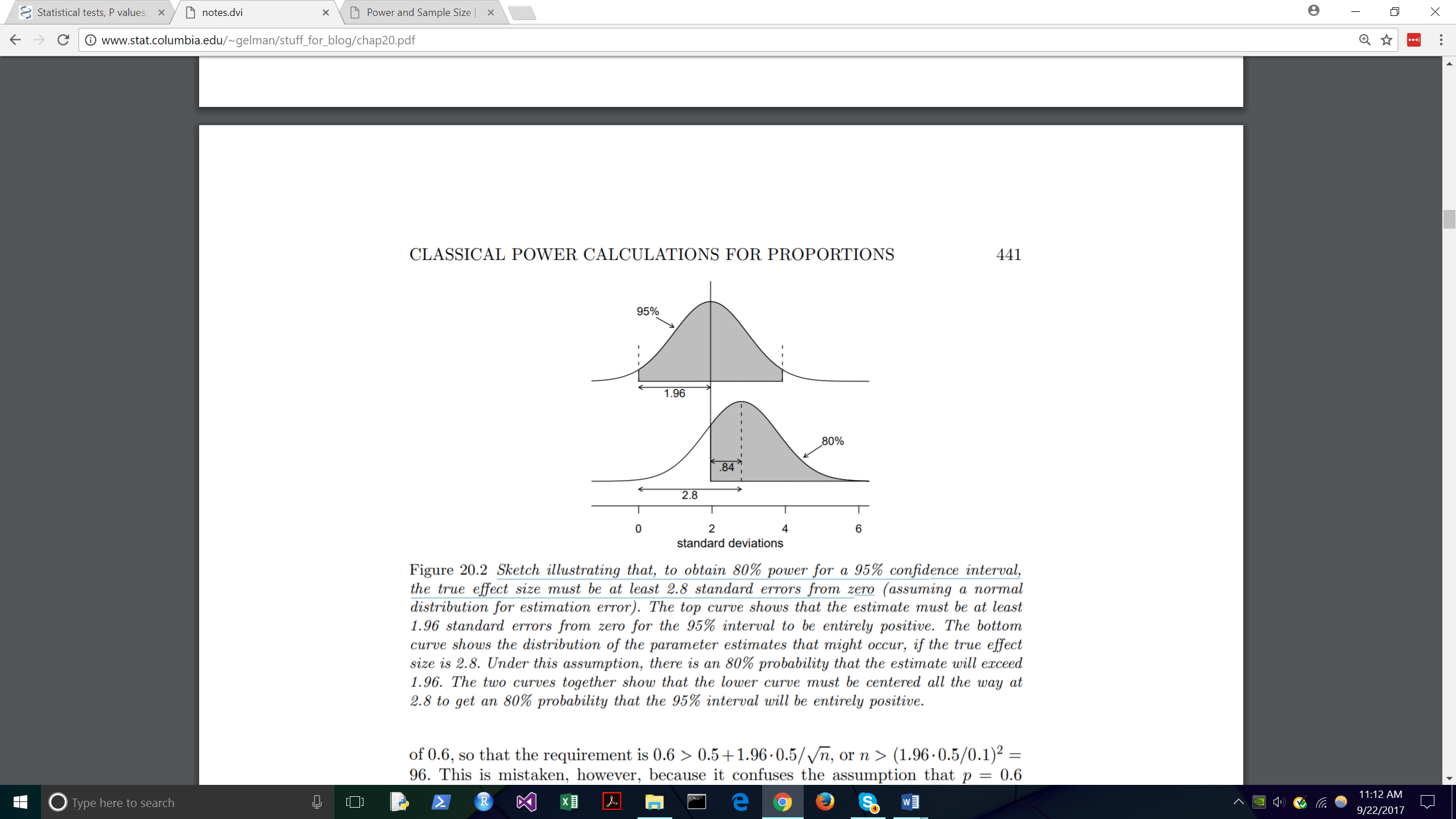
I like this page above.

<http://www.stat.columbia.edu/~gelman/stuff_for_blog/chap20.pdf>

The above has 20 pages on the topic. But here’s the section that initially catches my attention:

Sample size to achieve a specified probability of obtaining statistical significance.

Sketch illustrating that, to obtain 80% power for a 95% confidence interval, the true effect size must be at least 2.8 standard errors from zero



1. If we want 80% statistical power at 98% confidence interval…

* We target > 50% probability of having a mean of 1.96 standard deviations.
* It’s now a probability of the mean of the treatment.

To find the value of n such that exactly 80% of the estimates will be at least 1.96 standard errors from 0.5, we need:

0.5 + 1.96 s.e. = 0.6 − 0.84 s.e.

# Cross Validation in Detail

## CV Basics

Here’s what a basic cross validation will do. Think about linear regression, e.g. Result = x1A + x2B + x3C. We test this regression on 5 folds in a cross validation. You have five different cuts at getting the x1 – x3 weights.

You can train and test on the same dataset. It saves space, and potentially increases statistical power – if working with a small sample dataset.

There are LOTS of places where you can read about cross validation. Here, for example, is documentation from the sciki-learn site.

[https://scikit-learn.org/stable/modules/cross\_validation.html#](https://scikit-learn.org/stable/modules/cross_validation.html)

## Alternative: Nested CV / GridSearchCV

Nested CV is about manipulating hyper-parameters before manipulating parameters.

In machine learning, a hyperparameter is a parameter whose value is set before the learning process begins. By contrast, the values of other parameters are derived via training.

Think about it this way. A linear regression might find the correct values for x1, x2, and x3. But we have other inputs to these equations – e.g. the ‘learning rate’.

It is NOT Feature Engineering / selection of the correct features. It is NOT training the model. It is tuning the assumptions that go into the model – before you run the model.

Hyper-parameters by definition are input parameters which are necessarily required by an algorithm to learn from data.

# Bootstrapping

#### What it is (From Wikipedia)

The basic idea of bootstrapping is that inference about a population from sample data, (sample → population), can be modeled by *resampling* the sample data and performing inference about a sample from resampled data, (resampled → sample).

As the population is unknown, the true error in a sample statistic against its population value is unknown.

In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data, (resampled → sample), is measurable

#### Advantage

It is a straightforward way to derive estimates of [standard errors](https://en.wikipedia.org/wiki/Standard_error_(statistics)) and [confidence intervals](https://en.wikipedia.org/wiki/Confidence_intervals) for complex estimators

#### Disadvantage

It does not provide general finite-sample guarantees

#### Use Bootstrapping When:

1. The theoretical distribution of a statistic of interest is complicated or unknown
2. When the sample size is insufficient for straightforward statistical inference
3. When power calculations have to be performed, and a small pilot sample is available.

# Classifier: Support Vector Machines

# Some Basics / Terms and Purposes

## Accuracy, Precision, Recall, Confusion Matrix

### Accuracy: % of Population that’s properly classified

(True Positive + True Negative) / Total Population

### Precision: % of Positives that are actually positive

True Positive / (True Positive + False Positive)

## False Discovery Rate: % of Predictions that are False Positive

False Positive / (True Positive + False Positive)

### Recall: % of Positives that are flagged as positive

True Positive / (True Positive + False Negative)

### The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | True Condition = positive | True Condition = negative |
| Predicted  = Positive | True positive | False positive, Type I error |
| Predicted  = Negative | False negative, Type II error | True negative |

## ROC Curve

The ROC Curve (receiver operating characteristics, name from communications theory) tracks True Positive Rate vs False Positive Rate. Essentially – a plot where you have a very high % of the plot in area under the curve means you have a pretty good predictor.

* Henle: The ROC chart is the FIRST THING you should look at after finishing a classification.
* The ROC chart shows performance of the classifier across a RANGE of thresholds.
  + Confusion matrix is a snapshot of one threshold, e.g. 5%.

## Chi-Squared Test (vs Z-Test)

## A test of independence - determine if two variables are independent

* (Observed Frequency – Expected Frequency (Null, Independent) / Sample Size)

## A Test of Fit – (Pearson’s) Do Categories Fit a (particular) distribution?

* Calculate a statistics on this data. And we expect that this statistic has a chi-squared distribution, approximately. Sum (Observed – Expected), Squared. / Expected.

It might be: X^2 = (30-20)^2/20 + (14-20)^2/20 + (34-30)^2/30 + (45-40)^2/40 + (57-60)^2/60+(20-30)^2/30. = 11.44 – And we think that this # has a chi squared distribution.

So – what’s the probability that this is a more extreme result than the critical chi-squared value for a 5% value? Well – we start with six sums, which means we have five degrees of freedom. And look it up on the chart.

A 5% significance on 5 degrees of freedom is 11.07. Compare this to the sample, 11.44 here. This is a cause to reject the null hypothesis.

# Z-Test vs T-Test vs Chi Squared vs Fisher’s

1. z-test. A z-test assumes that our observations are independently drawn from a Normal distribution with unknown mean and known variance. A z-test is used primarily when we have quantitative data. (i.e. weights of rodents, ages of individuals, systolic blood pressure, etc.) However, z-tests can also be used when interested in proportions. (i.e. the proportion of people who get at least eight hours of sleep, etc.)
2. t-test. A t-test assumes that our observations are independently drawn from a Normal distribution with unknown mean and unknown variance. Note that with a t-test, we do not know the population variance. This is far more common than knowing the population variance, so a t-test is generally more appropriate than a zz-test, but practically there will be little difference between the two if sample sizes are large.

With zz- and tt-tests, your alternative hypothesis will be that your population mean (or population proportion) of one group is either not equal, less than, or greater than the population mean (or proportion) or the other group. This will depend on the type of analysis you seek to do, but your null and alternative hypotheses directly compare the means/proportions from the two groups.

1. Chi-squared test. Whereas zz- and tt-tests concern quantitative data (or proportions in the case of zz), **chi-squared tests are appropriate for qualitative data**. Again, the assumption is that observations are independent of one another. In this case, you aren't seeking a particular relationship. Your null hypothesis is that no relationship exists between variable one and variable two. Your alternative hypothesis is that a relationship does exist. This doesn't give you specifics as to how this relationship exists (i.e. In which direction does the relationship go) but it will provide evidence that a relationship does (or does not) exist between your independent variable and your groups.
2. Fisher's exact test. One drawback to the chi-squared test is that it is asymptotic. This means that the pp-value is accurate for very large sample sizes. However, if your sample sizes are small, then the pp-value may not be quite accurate. As such, Fisher's exact test allows you to exactly calculate the pp-value of your data and not rely on approximations that will be poor if your sample sizes are small.

If you are looking for a specific effect from your A/B test (for example, my B group has higher test scores), then I would opt for a z-test or t-test, pending sample size and the knowledge of the population variance.

If you want to show that a relationship merely exists (for example, my A group and B group are different based on the independent variable but I don't care which group has higher scores), then the chi-squared or Fisher's exact test is appropriate, depending on sample size.

## What are Degrees of Freedom? – Values In the Math That Can Vary

Wikipedia:

In the strictest definition, degrees of freedom are the number or ways that a final statistic can vary.

* A common way to think of degrees of freedom is as the number of independent pieces of information available to estimate another piece of information
* More concretely, the number of degrees of freedom is the number of independent observations in a sample of data that are available to estimate a parameter
  + of the population from which that sample is drawn.

In Probability Distributions

Several commonly encountered statistical distributions (Student's t, Chi-Squared, F) have parameters that are commonly referred to as degrees of freedom.

* This terminology simply reflects that in many applications where these distributions occur, the parameter corresponds to the degrees of freedom of an underlying random vector